Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer

V. G. Kulkarni, L. Gün, *Senior Member, IEEE,* and P. F. Chimento

Abstract—**We consider a traffic model where a single source generates traffic having J ($J \geq 2$) quality of service (QoS) classes. QoS in this case is described by a cell loss probability objective ϵ_j for QoS class j. We assume that $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_J$, in other words, class J has the most stringent QoS requirements and class 1 the least. The traffic from K such independent heterogeneous Markov-modulated fluid sources is multiplexed into a single buffer of size B. There are $J-1$ thresholds $\{B_j, 1 \leq j \leq J-1\}$ such that $0 < B_1 < B_2 < \cdots < B_{J-1} < B$. Let $B_0 = 0$ and $B_J = B$. If the buffer content is in $(B_{j-1}, B_j), 1 \leq j \leq J$ only traffic of class index j or above is accepted and all other traffic is rejected. For this system of K sources we define an effective bandwidth vector of size J such that QoS requirements for all classes are satisfied if each component of the vector is less than the channel capacity. We propose several bandwidth vectors that can be computed for each source separately. Numerical studies are reported on the efficacy of these bandwidth vectors.**

I. Introduction

THE emerging high speed networks achieve efficiency and higher resource utilization by using statistical multiplexing. Under statistical multiplexing the traffic from several sources gets superimposed onto a single buffer and is transmitted to the network in a first-come-first-served fashion. In the high-speed networks using asynchronous transfer mode (ATM), each source is assured a given quality of service (QoS), and the admission control scheme is designed to assure that a source is admitted into the network only if there is sufficient capacity in the network to guarantee the QoS.

Stochastic fluid models and the theory of large deviations have been used in the literature to define what is called *effective bandwidth* or *equivalent capacity* for each source. The effective bandwidth is a number that depends on the source characteristic, the QoS requirement of the source and the buffer size. The most useful property of the effective bandwidth is the following: If the sum of the effective bandwidths of all the sources multiplexed onto a buffer is less than the speed of the channel removing data from the buffer then (in certain asymptotic sense) the QoS requirements of all the sources are satisfied. This is what makes the concept of effective bandwidths very useful in admission control. Several authors have studied this concept from different angles [3], [6]–[8], [9], [11].

The above method of admission control indeed works very satisfactorily as long as the QoS reuirements of all the sources are the same or at least close. In practice all the sources get the QoS of the most stringent source, since they all get the same QoS. This manifests itself in less than optimal resource allocation. This is a major drawback of this method.

When the source QoS requirements are varied, or when every source generates traffic belonging to multiple classes with varying QoS requirements (such as MPEG-2 multilayered video, or sources policed by leaky bucket regulators), the above call admission procedure needs to be changed. In particular, when a source, such as a multilayered video codec, transmits, the information from different layers that emanate from that single source may have very different QoS requirements. So, for example, some signals may have only enhancement effects on the final image, or enhancements of sound quality, and so can be transmitted at a much lower QoS than signals having to do with the more fundamental quality of the way the information is presented [13].

One method is to classify the traffic into multiple classes according to their QoS requirements and to employ a separate buffer for each class of traffic. Then the effective bandwidth methodology can be applied to each buffer separately. This, however, has several drawbacks: First, we need a scheduler to schedule the transmission from the various buffers to ensure first in first out discipline for each connection (source). Second, the buffer utilization is inferior. Third, if the cells of different priorities emanate from a single source (that is, they are part of a single stream), then having seperate buffers for differing QoS requirements forces resequencing at the destination, which is highly undesirable. Fourth, having seperate logical buffers for each QoS requirement complicates the implementation.

To overcome the above disadvantages one can consider a shared buffer scheme. Under this scheme the traffic is classified as before. Suppose there are J classes, indexed 1, 2, $\cdots$, J. We assume that class J has the most stringent QoS and class 1 has the least stringent one. The buffer is of size B. There are $J-1$ thresholds $\{B_j, 1 \leq j \leq J-1\}$ such that $0 < B_1 < B_2 < \cdots < B_{J-1} < B$. For convenience, define $B_0 = 0$ and $B_J = B$. If the buffer content is in the interval $(B_{j-1}, B_j), (1 \leq j \leq J)$ only traffic of class index

Manuscript received September 30, 1994; revised April 1, 1995. This work was supported in part by NSF Grant NCR-9406823. This paper was presented in part at the INFORMS Telecom Conference, Boca Raton, FL, March 1995.

V. G. Kulkarni was a visitor at IBM High Bandwidth Networking, Network Analysis Center, Research Triangle Park, NC 27709 USA. He is now with the Department of Operations Research, University of North Carolina, Chapel Hill, NC 27599 USA.

L. Gün was with IBM High Bandwidth Networking, Network Analysis Center, Research Triangle Park, NC 27709 USA. He is now with Motorola Codex, Mansfield, MA 02048 USA.

P. F. Chimento is on leave with IBM High Bandwidth Networking, Network Analysis Center, Research Triangle Park, NC 27709 USA. He is now with Tele-Informatics and Open Systems, University of Twente, Enschede, The Netherlands.

IEEE Log Number 9412653.

j or above is accepted and all other traffic is rejected. The admission policy is more complicated when the buffer content is exactly at a threshold and will be stated precisely in the next section.

The above scheme of differentiating traffic from different traffic classes is called a buffer sharing scheme. It is intuitively clear that the above mechanism would provide highest QoS to class J and lowest QoS to class 1 traffic. Furthermore, the scheme maintains the first-in-first-out ordering among all accepted traffic and eliminates the need for schedulers. It is also easy to analyse [3], [13] and is shown to be optimal within a large class of service policies [2], [12], [14]. For these reasons this scheme is very attractive.

In this paper we analyse this scheme and show how the concept of effective bandwidths can be extended to this configuration. The aim is construct a vector of size J (called the effective bandwidth vector) for each source such that the QoS requirements are satisfied for all classes if the each component of the sum of the effective bandwidth vectors is less than the channel speed.

The purpose of constructing such a vector is to try to reduce the bandwidth requirements of the source stream as a whole. So, for example, if no priority differentiation is made, then all the cells from that particular source must be treated as though they required the most stringent QoS. Intuitively, one would expect the bandwidth requirements to be less when there is priority differentiation. For a given number of sources and for a given trunk capacity, bandwidth savings can be defined as the difference between the total effective bandwidth required by the sources with no differentiation and that required by the sources with two or more levels of priority.

The paper is organized as follows: The model is described in detail, and the bulk of the notation is introduced in Section II. Section III contains the relevant results from the descriptive analysis of this model. This section essentially restates the results of [3] using our notation. In Section IV we present the asymptotic analysis of the loss probability (QoS). The asymptotic region is precisely defined in this section. It also states a sufficient condition for the priority differentation to produce savings. Using the results of this section, we construct an effective bandwidth vector for the system of K sources in Section V. Unfortunately, this bandwidth vector is not additive, i.e., it cannot be written as a sum of K effective bandwidth vectors, one for each source. Hence we study several possible candidates for effective bandwidth vectors for a single source. The most logical candidate, unfortunately, provides only an approximation, and not a bound. Hence, we propose several candidates that can be proved to provide bounds, although loose. In Section VI we discuss numerical algorithms to compute the effective bandwidth vectors. We illustrate the concepts in Section VII by numerical examples using a voice-multiplexing example.

II. The Model

In this section we describe the precise model of multiplexing K Markov modulated fluid sources, each producing fluid belonging to J different classes onto a single buffer of size B that is serviced by a channel of capacity c. Let $Z_k(t)$ be the state of the kth source (i.e., the state of the external environment that controls the traffic stream generated by the kth source) at time t. $\{Z_k(t), t \geq 0\}$ is assumed to be an irreducible continuous time Markov chain (CTMC) with state space $\mathcal{N}_k = \{1, 2, \cdots, \mathcal{N}_k\}$ and generator matrix $G_k, (1 \leq k \leq K)$. When the kth source is in state i, it generates class j fluid at rate $\lambda_{k,i}^j$. As explained in the introduction, the admission policy is based upon a space reservation scheme, using thresholds $\{B_j, 1 \leq j \leq J-1\}$.

Let $X(t)$ be the amount of fluid (of all classes) in the buffer at time t. When $B_{j-1} < X(t) < B_j$, only fluid belonging to classes $\{j, j+1, \cdots, J\}$ is admitted into the buffer. Thus, when $0 \leq X(t) < B_1$, fluid of all classes is admitted, while when $B_{J-1} < X(t) < B_J$, only fluid of class J is admitted. At threshold B_j fluid of class j is admitted at such rate that the buffer content does not rise above B_j. To make this precise, we give an expression for $I_j(t)$, the rate at which the fluid of class j is admitted into the buffer at time t. First, we need the following notation

$$\Lambda_{k,i}^j = \sum_{r=j}^{J} \lambda_{k,i}^r. \tag{1}$$

With this we can write (2), seen at the bottom of this page, for $1 \leq j \leq J$. We use $(x)^+ = \max(0, x)$ and $(x)^- = \min(0, x)$. Now, let $\Phi(B_j)$ be the long-run probability that the buffer content is above the threshold B_j. We take this probability to be the surrogate for the loss fraction for fluid of class j. The QoS guarantee is that this loss fraction is kept below a prespecified quantity ϵ_j. Obviously, $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_J > 0$. Thus, using the surrogate, the QoS is satisfied for all the classes of traffic if

$$\Phi(B_j) < \epsilon_j, \qquad \text{for all } 1 \leq j \leq J. \tag{3}$$

III. The Analysis

Let $Z(t) = (Z_1(t), Z_2(t), \cdots, Z_K(t))$ be the state vector of the K sources. It is clear that $\{(X(t), Z(t)), t \geq 0\}$ is a Markov process with piecewise deterministic paths. The

$$I_j(t) = \begin{cases} \sum_{k=1}^{K} \lambda_{k,Z_k(t)}^j, & \text{if } X(t) < B_j \\ \min\{\sum_{k=1}^{K} \lambda_{k,Z_k(t)}^j, (c - \sum_{k=1}^{K} \Lambda_{k,Z_k(t)}^{j+1})^+\}, & \text{if } X(t) = B_j \\ 0, & \text{if } X(t) > B_j \end{cases} \tag{2}$$

dynamics of the process is given by

$$\frac{dX(t)}{dt} = \begin{cases} (\sum_{j=1}^{J} I_j(t) - c)^+, & \text{if } X(t) = 0 \\ \sum_{j=1}^{J} I_j(t) - c, & \text{if } 0 < X(t) < B \\ (\sum_{j=1}^{J} I_j(t) - c)^-, & \text{if } X(t) = B. \end{cases} \quad (4)$$

Now, $\{Z(t), t \geq 0\}$ is a CTMC with state-space $\mathcal{N} = \mathcal{N}_1 \times \mathcal{N}_2 \times \cdots \times \mathcal{N}_K$ where $\mathcal{N}_k = \{1, 2, \cdots N_k\}$ is the state-space of $\{Z_k(t), t \geq 0\}$. The generator of $\{Z(t), t \geq 0\}$ is given by

$$G = G_1 \oplus G_2 \oplus \cdots \oplus G_K \quad (5)$$

where $\oplus$ represents the Kronecker sum. Let

$$p_k(i) = \lim_{t \to \infty} P\{Z_k(t) = i\}, \qquad (i \in \mathcal{N}_k, k = 1, 2, \cdots, K) \quad (6)$$

$$p(n) = \lim_{t \to \infty} P\{Z(t) = n\}, \qquad (n \in \mathcal{N}) \quad (7)$$

be the limiting distribution of Z. Now let

$$\pi(x, n) = \lim_{t \to \infty} P\{X(t) \geq x, Z(t) = n\}, \qquad (x \geq 0, n \in \mathcal{N}) \quad (8)$$

and

$$\pi(x) = [\pi(x, n)]_{n \in \mathcal{N}}. \quad (9)$$

For $1 \leq j \leq J$, we use the notation

$$\pi^j(x) = \pi(x), \qquad B_{j-1} < x < B_j. \quad (10)$$

Next, we state the differential equations satisfied by $\pi^j(x)$. First, we need the following notation

$$\Lambda_k^j = \text{diag}(\Lambda_{k,1}^j, \Lambda_{k,2}^j, \cdots, \Lambda_{k,N_k}^j), \qquad 1 \leq j \leq J \quad (11)$$

$$\Lambda^j = \Lambda_1^j \oplus \Lambda_2^j \oplus \cdots \oplus \Lambda_K^j, \qquad 1 \leq j \leq J \quad (12)$$

$$D_j = \Lambda^j - cI, \qquad 1 \leq j \leq J. \quad (13)$$

The $\{\pi^j(x), 1 \leq j \leq J\}$ satisfy the following differential equations

$$\frac{d}{dx}\pi^j(x) D^j = \pi^j(x) G, \qquad B_{j-1} < x < B_j. \quad (14)$$

Next, we state the boundary conditions. For $1 \leq j \leq J$, let

$$\mathcal{N}_+^j = \{n \in \mathcal{N} : D^j(n, n) > 0\} \quad (15)$$

$$\mathcal{N}_-^j = \{n \in \mathcal{N} : D^j(n, n) < 0\}. \quad (16)$$

For simplicity we assume that $\{n \in \mathcal{N} : D^j(n, n) = 0\} = \emptyset$. (The final results hold true even when this assumption is not satisfied.) Note that

$$\mathcal{N}_+^j \cup \mathcal{N}_-^j = \mathcal{N}, \qquad 1 \leq j \leq J \quad (17)$$

$$\mathcal{N}_-^1 \subseteq \mathcal{N}_-^2 \subseteq \cdots \subseteq \mathcal{N}_-^J \quad (18)$$

$$\mathcal{N}_+^1 \supseteq \mathcal{N}_+^2 \supseteq \cdots \supseteq \mathcal{N}_+^J. \quad (19)$$

With this notation we can write the boundary conditions as follows (See Elwalid and Mitra [3] for their derivation.)

$$\pi^1(0, n) = p(n), \qquad n \in \mathcal{N}_+^1 \quad (20)$$

$$\pi^j(B_j-, n) = \pi^{j+1}(B_j+, n), \qquad n \in \mathcal{N}_-^j \cup \mathcal{N}_+^{j+1} \quad (21)$$

$$\pi^J(B_J-, n) = 0, \qquad n \in \mathcal{N}_-^J. \quad (22)$$

The above boundary conditions are intuitive. Equation (21), for example, says that there is no mass at (B_j, n) in steady state if the drift on either side of B_j is away from B_j. The spectral solution to (14) is given by

$$\pi^j(x) = \sum_r a_r^j \phi_r^j e^{\eta_r^j x}, \qquad 1 \leq j \leq J \quad (23)$$

where each (η_r^j, ϕ_r^j) pair is an (eigenvalue, eigenvector) pair satisfying

$$\eta \phi D^j = \phi G. \quad (24)$$

Elwalid and Mitra in [4] show how the (eigenvalue, eigenvector) problem of the above equation involving the large G matrix can be reduced to a coupled (eigenvalue, eigenvector) problem invoving the smaller matrices G_k. We briefly state the main result. For $j, 1 \leq j \leq J$, and $k, 1 \leq k \leq K$, define

$$A_k^j(\eta) = \Lambda_k^j - \frac{1}{\eta} G_k. \quad (25)$$

Theorem 3.1–Elwalid-Mitra: Let $1 \leq j \leq J$ be fixed. i) A pair (η, ϕ) satisfies (24) if and only if the following equations hold

$$g_k^j(\eta) \phi_k^j = \phi_k^j A_k^j(\eta), \qquad 1 \leq k \leq K \quad (26)$$

$$\sum_{k=1}^{K} g_k^j(\eta) = c \quad (27)$$

$$\phi = \phi_1^j \otimes \phi_2^j \otimes \cdots \otimes \phi_K^j. \quad (28)$$

ii) For $\eta < 0$ the solution $g_k^j(\eta)$ to (26) with the maximum real part is a simple real solution, called the maximal real eigenvalue, denoted by $g_k^{j*}(\eta)$ and it decreases monotonically from $\Lambda_k^{j,max} (= \max_{i=1,\cdots,N_k} \Lambda_{k,i}^j)$ to $\Lambda_k^{j,mean} (= \sum_{i=1}^{N_k} p_k(i) \Lambda_{k,i}^j)$ as η increases from $-\infty$ to 0. iii) For $\Lambda^{j,mean} (= \sum_{k=1}^{K} \Lambda_k^{j,mean}) < c < \Lambda^{j,max} (= \sum_{k=1}^{K} \Lambda_k^{j,max})$, the dominant eigenvalue η^{j*}, is given by the unique solution in $(-\infty, 0)$ to

$$\sum_{k=1}^{K} g_k^{j*}(\eta^{j*}) = c. \quad (29)$$

Furthermore, η^{j*} is a monotonic, strictly decreasing function of $c \in (\Lambda^{j,\text{mean}}, \Lambda^{j,\text{max}})$.

We consider the case of distinct eigenvalues, in which case we have $|\mathcal{N}|$ eigenvalues. The unknowns $\{a_r^j, 1 \leq j \leq J, r = 1, 2, \cdots, |\mathcal{N}|\}$ are obtained by using (20)–(22) which provide the correct number of equations for them. Now, for $1 \leq j \leq J$

$$\Phi(B_j) = P\{\text{BufferContent} \geq \text{B}_\text{j}\} \quad (30)$$

$$= \sum_{n \in \mathcal{N}} \pi^j(B_j, n). \quad (31)$$

Thus, the QoS requirements can be roughly written as

$$\Phi(B_j) \leq \epsilon_j, \qquad 1 \leq j \leq J. \quad (32)$$

Now that we have a way of computing the QoS requirements, we next study the asymptotic behavior of $\Phi(B_j)$ as buffer size gets large and the QoS requirements get more stringent.

IV. Asymptotics

In this section we study the following asymptotic case

$$B_j \to \infty, \qquad 1 \le j \le J \tag{33}$$

$$\text{and} \quad \epsilon_j \to 0, \qquad 1 \le j \le J \tag{34}$$

$$\text{so that} \quad \frac{B_j - B_{j-1}}{B_J} \to b_j, \qquad 1 \le j \le J \tag{35}$$

$$\text{and} \quad \frac{log\epsilon_j}{B_j} \to \gamma_j, \qquad 1 \le j \le J \tag{36}$$

where $b_j > 0, \sum_{j=1}^{J} b_j = 1$ and $-\infty < \gamma_j < 0$. Note that the condition $b_j > 0$ implies that all threshold seperations $B_j - B_{j-1}$ tend to infinity in the asymptotic region. The next theorem gives the asymptotic behavior of $\Phi(B_j)$.

Theorem 4.1: In the asymptotic case described by (33) and (35) we have

$$\Phi(B_j) = \exp\{\sum_{r=1}^{j} \eta^{r*}(B_r - B_{r-1})\}\{\text{constant} + o(1)\}, \qquad 1 \le j \le J. \tag{37}$$

Proof: The proof is a tedious exercise in matrix algebra. We start by introducing the required notation

$$\mathcal{R}_+^j = \{r : \mathrm{Re}(\eta_r^j) \ge 0\} \tag{38}$$

$$\mathcal{R}_-^j = \{r : \mathrm{Re}(\eta_r^j) < 0\} \tag{39}$$

$$a^j = [a_r^j]_{r=1,\cdots,\mathcal{N}} \tag{40}$$

$$a_+^j = [a_r^j : r \in \mathcal{R}_+^j] \tag{41}$$

$$a_-^j = [a_r^j : r \in \mathcal{R}_-^j] \tag{42}$$

$$\phi^j(A) = [\phi_r^j(n)]_{n\in A}, \text{ for } A \subseteq \mathcal{N} \tag{43}$$

$$\phi_+^j(A) = [\phi_r^j(n)]_{n\in A, r\in\mathcal{R}_+^j} \tag{44}$$

$$\phi_-^j(A) = [\phi_r^j(n)]_{n\in A, r\in\mathcal{R}_-^j} \tag{45}$$

$$p(A) = [p(n)]_{n\in A} \tag{46}$$

$$E^j(x) = \mathrm{Diag}[e^{\eta_r^j x}]_{r=1,\cdots,\mathcal{N}} \tag{47}$$

$$E_+^j(x) = \mathrm{Diag}[e^{\eta_r^j x}]_{r\in\mathcal{R}_+^j} \tag{48}$$

$$E_-^j(x) = \mathrm{Diag}[e^{\eta_r^j x}]_{r\in\mathcal{R}_-^j}. \tag{49}$$

With the above notation the QoS requirements of (32) can be written as

$$\Phi(B_j) = a^j E^j(B_j)\phi^j(\mathcal{N})1 \le \epsilon_j, \qquad \text{for } 1 \le j \le J \tag{50}$$

where 1 is a column vector of dimension $|\mathcal{N}|$. Using the above notation we shall prove the theorem for the case $J = 2$. The general case follows in a similar fashion. The boundary conditions in (20)–(22) can be written as

$$a^1\phi^1(\mathcal{N}_+^1) = p(\mathcal{N}_+^1) \tag{51}$$

$$a^1E^1(B_1)\phi^1(\mathcal{N}_-^1) = a^2E^2(B_1)\phi^2(\mathcal{N}_-^1) \tag{52}$$

$$a^1E^1(B_1)\phi^1(\mathcal{N}_+^2) = a^2E^2(B_1)\phi^2(\mathcal{N}_+^2) \tag{53}$$

$$a^2E^2(B_2)\phi^2(\mathcal{N}_-^2) = 0. \tag{54}$$

Equation (54) yields

$$a_+^2E_+^2\phi_+^2(\mathcal{N}_-^2) + a_-^2E_-^2\phi_-^2(\mathcal{N}_-^2) = 0 \tag{55}$$

which yields

$$a_+^2E_+^2(B_2) = -a_-^2E_-^2(B_2)\phi_-^2(\mathcal{N}_-^2)\phi(\mathcal{N}_-^2)^{-1}. \tag{56}$$

Next, (53) yields

$$a^1E^1(B_1)\phi(\mathcal{N}_+^1)$$

$$= a_+^2E_+^2(B_1)\phi_+^2(\mathcal{N}_+^2) + a_-^2E_-^2(B_1)\phi_-^2(\mathcal{N}_+^2) \tag{57}$$

$$= a_+^2E_+^2(B_2)E_+^2(B_1 - B_2)\phi_+^2(\mathcal{N}_+^2) + a_-^2E_-^2(B_1)\phi_-^2(\mathcal{N}_+^2) \tag{58}$$

$$= a_-^2E_-^2(B_2)[-\phi_-^2(\mathcal{N}_-^2)\phi_+^2(\mathcal{N}_-^2)^{-1}E_+^2(B_1 - B_2)\phi_+^2(\mathcal{N}_+^2) + E_-^2(B_1 - B_2)\phi_-^2(\mathcal{N}_+^2)]. \tag{59}$$

The last equation follows by substituting (56) in (58). Now, in the asymptotic region under consideration, $E_+^2(B_1 - B_2)$ remains bounded above by 1, while $E_-^2(B_1 - B_2)$ goes to ∞. Hence the first term in the square brackets in (59) can be ignored in the asymptotic region. Hence, we get

$$a_-^2E_-^2(B_2) = a^1E^1(B_1)[\cdot]E_-^2(B_2 - B_1) \tag{60}$$

where $[\cdot]$ represents a matrix that does not dependent on B_1 or B_2. Using this convention we get

$$a^2E^2(B_2) = a_+^2E_+^2(B_2) + a_-^2E_-^2(B_2) \tag{61}$$

$$= a_-^2E_-^2(B_2)[\cdot] \tag{62}$$

$$= a^1E^1(B_1)[\cdot]E_-^2(B_2 - B_1)[\cdot]. \tag{63}$$

A similar analysis shows that

$$a^1E^1(B_1) = p(\mathcal{N}_+^1)[.]E_-^1(B_1). \tag{64}$$

Combining (63)–(64) we get

$$a^2E^2(B_2) = [.]E_-^1(B_1)[.]E_-^2(B_2 - B_1)[.]. \tag{65}$$

Now, (64) shows that, in the asymptotic region, $\Phi(B_1)$ is a linear combination of terms of the type $e^{\eta_r^1 B_1}, r \in \mathcal{R}_-^1$, and (65) shows that $\Phi(B_2)$ is a linear combination of terms of the type $e^{\eta_r^1 B_1 + \eta_s^2(B_2 - B_1)}, r \in \mathcal{R}_-^1, s \in \mathcal{R}_-^2$. Using the dominant terms corresponding to η^{1*} and η^{2*} we get the theorem. □

Since there can be a nonzero probability mass at B_j, it is important to realize that $\Phi(B_j)$ in the above theorem is in fact $\Phi(B_j+)$, that is, it includes that mass. The next theorem gives a sufficient condition to check if the QoS criterion is satisfied for the traffic of all classes.

Theorem 4.2: The QoS criterion of (32) is satisfied in the asymptotic region of (33)–(36) if there exist negative numbers $\{\gamma_{jr}, 1 \le j \le J, 1 \le r \le j\}$ such that

$$\sum_{r=1}^{j} b_r\gamma_{jr} = \gamma_j \sum_{r=1}^{j} b_r \tag{66}$$

and

$$\sum_{k=1}^{K} g_k^{r*}(\gamma_{jr}) < c, \qquad 1 \le r \le j. \tag{67}$$

Proof: From Theorem 2, in the asymptotic region of (33)–(36) we can write

$$\frac{\Phi(B_j)}{\epsilon_j} = \exp\{\delta_j B_J\}\{\text{constant} + o(1)\}, \qquad 1 \le j \le J. \tag{68}$$

where

$$\delta_j = \sum_{r=1}^{j} \eta^{r*} b_r - \gamma_j \sum_{r=1}^{j} b_r. \tag{69}$$

Now, $\Phi(B_j)/\epsilon_j \to 0$ if $\delta_j < 0$ and $\Phi(B_j)/\epsilon_j \to \infty$ if $\delta_j > 0$ in the asymptotic region under consideration. It can be easily seen that $\delta_j < 0$ if and only if there exists a set of negative numbers $\{\gamma_{jr}, 1 \le r \le j\}$ satisfying (66) and

$$\eta^{r*} < \gamma_{jr}, \qquad 1 \le r \le j. \tag{70}$$

Now, from EM it follows that $\eta^{r*} < \gamma_{jr}$ if $\sum_{k=1}^{K} g_k^{r*}(\gamma_{jr}) < c$. Hence the theorem follows. □

One basic question in priority traffic is: Will the creation of several priority classes always result in savings? The next theorem shows that the answer is a conditional "yes."

Theorem 4.3: The priority differentiation results in savings if

$$\gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_J. \tag{71}$$

Proof: We shall give a very restricted proof, but it will provide an idea behind the general result. Consider two systems, each with K sources, that are identical in all respects except that priority one traffic of system one is treated as priority two traffic in system two. Using an additional superscript l to denote the system index $l = 1, 2$, we have

$$\lambda_{k,i}^{j,2} = \begin{cases} 0, & \text{for } j = 1 \\ \lambda_{k,i}^{1,1} + \lambda_{k,i}^{2,1}, & \text{for } j = 2 \\ \lambda_{k,i}^{j,1}, & \text{for } j = 3, \cdots, J. \end{cases} \tag{72}$$

This implies that

$$g_k^{j*,2}(\eta) = \begin{cases} g_k^{1*,1}(\eta), & \text{for } j = 1, 2 \\ g_k^{j*,1}(\eta), & \text{for } j = 3, \cdots, J. \end{cases} \tag{73}$$

Now for system two, the sufficient conditions of Theorem 3 need to be satisfied only for $j = 2, \cdots, J$. Suppose this is the case. The condition in (67) for $j = 2$, when applied to system two, becomes

$$\sum_{k=1}^{K} g_k^{2*,1}(\gamma_{21}) < c \tag{74}$$

$$\sum_{k=1}^{K} g_k^{2*,2}(\gamma_{22}) < c. \tag{75}$$

Using (73) and the fact that $g_k^{1*,1}(\eta)$ is a decreasing function of η the above equations can be written as

$$\sum_{k=1}^{K} g_k^{1*,1}(\min\{\gamma_{21}, \gamma_{22}\}) < c. \tag{76}$$

Since the maximum value of $\min\{\gamma_{21}, \gamma_{22}\}$ is γ_2 we see that the above equation implies

$$\sum_{k=1}^{K} g_k^{1*,1}(\gamma_2) < c. \tag{77}$$

Now, if $\gamma_2 \le \gamma_1$, then (77) implies

$$\sum_{k=1}^{K} g_k^{1*,1}(\gamma_1) < c. \tag{78}$$

Furthermore, $g_k^{2*,1}(\eta) \le g_k^{1*,1}(\eta)$ implies

$$\sum_{k=1}^{K} g_k^{2*,1}(\gamma_2) < c. \tag{79}$$

Thus, condition (67) is satisfied for system one for $j = 1$ with $\gamma_{11} = \gamma_1$ and for $j = 2$ with $\gamma_{21} = \gamma_{22} = \gamma_2$. For $j = 3, \cdots, J$ the conditions in (67) are the same for both the system. Thus, if the sufficient conditions of Theorem 3 are satisfied for system two, then they are satisfied for system one, provided $\gamma_2 \le \gamma_1$. This proves the result in this case. The general result follows in a similar fashion. □

Note that the condition in (71) implies that the ratio B_j/B_{j-1} must be larger than $\log(\epsilon_j)/\log(\epsilon_{j-1})$ in order to achieve savings through priority differentiation. In fact, when $\log(\epsilon_j)/B_j$ is a constant, (independent of j) one can see that the bandwidth requirements of a source with priority differentiation is the same as the one which treats all traffic as the highest priority traffic. This analytic insight is itself a benefit of this analysis.

We note here that in [1], Choudhury, Lucantoni, and Whitt discuss the notion of effective bandwidth and point out some of its drawbacks. In general, an effective bandwidth approximation based only on large buffer asymptotics can significantly overestimate or underestimate the number of sources that can be multiplexed on a trunk. If the sources are more "bursty" than Poisson, then effective bandwidth tends to underestimate the number of sources that can be multiplexed, and if the sources are less "bursty" then the number of sources are overestimated. This argues that effective bandwidth results should be used carefully, perhaps combined with other procedures as in [5].

V. Effective Bandwidth Vectors

The sufficient condition of Theorem 3 can be written in the following form

$$c^{j*}(K) < c, \quad 1 \le j \le J, \tag{80}$$

$$\text{where } c^{j*}(K) = \min_{(\gamma_{j1}, \cdots, \gamma_{jj}) \in \Delta_j} \max_{r=1,\cdots,j} \{\sum_{k=1}^{K} g_k^{r*}(\gamma_{jr})\}, \quad 1 \le j \le J, \tag{81}$$

$$\text{where } \Delta_j = \{(x_1, \cdots, x_j) : x_r < 0, 1 \le r \le j, \sum_{r=1}^{j} x_r b_r = \gamma_j \sum_{r=1}^{j} b_r\}. \tag{82}$$

This leads to the following natural definition:

Definition: The vector

$$c^*(K) = (c^{1*}(K), c^{2*}(K), \cdots, c^{J*}(K)) \tag{83}$$

is called the *effective bandwidth vector* for the system of K sources multiplexed onto a single buffer. However, it is clear from the definition that the effective bandwidth vector of the K-source system cannot be represented as a sum of K effective bandwidth vectors, each associated with a single source. In the next section we study one candidate for effective bandwidth vectors of a single source, which will serve as an approximation.

An Approximate Bandwidth Vector

Here we consider the case of independent and identical sources. Thus the generator matrix of the external environment driving each source is G, and the rate at which a source produces traffic of class j in state i is λ_i^j. Thus the maximal real eigenvalues are the same for all the sources, i.e. $g_k^{j*}(\eta) = g^{j*}(\eta)$. Then we get

$$c^{j*}(K) = \min_{(\gamma_{j1},\cdots,\gamma_{jj})\in\Delta_j} \max_{r=1,\cdots,j} \{\sum_{k=1}^{K} g_k^{r*}(\gamma_{jr})\} \tag{84}$$

$$= Kc^{j*} \tag{85}$$

$$\text{where } c^{j*} = \min_{(\gamma_{j1},\cdots,\gamma_{jj})\in\Delta_j} \max_{r=1,\cdots,j} \{g^{r*}(\gamma_{jr})\}. \tag{86}$$

Thus, we can define

$$c^* = (c^{1*}, c^{2*}, \cdots, c^{J*}) \tag{87}$$

as the effective bandwidth vector of a single source. (This is a direct generalization of the effective bandwidth vector defined in [10]). Thus as long as the sources are identical and independent, the effective bandwidth vector of the K-source system is the sum of the effective bandwidth vectors of the individual sources, i.e.

$$c^*(K) = Kc^*. \tag{88}$$

This motivates us to study the effective bandwidth vector defined by (87) as a candidate even if the sources are distinct.

Now, consider the K distinct sources as described in Section II. For the kth source, define the effective bandwidth vector as $c_k^* = (c_k^{1*}, c_k^{2*}, \cdots, c_k^{J*})$ where

$$c_k^{j*} = \min_{(\gamma_{j1},\cdots,\gamma_{jj})\in\Delta_j} \max_{r=1,\cdots,j} \{g_k^{r*}(\gamma_{jr})\}. \tag{89}$$

The question is, how does $\sum_{k=1}^{K} c_k^*$ compare with the system bandwidth vector $c^*(K)$ defined by (81)? Unfortunately, the sum provides neither an upper bound nor a lower bound. It only provides an approximation. This is because moving the sum in (81) outside the max operator increases the right-hand side, while further moving it outside the min operator decreases the right-hand side.

The next theorem gives an important result. Let $\gamma_k^{jr*}, 1 \le r \le j$ be the point where the minimum on the right-hand side of (89) is achieved. Also let $\gamma^{jr*}, 1 \le r \le j$ be the point where the minimum on the right-hand side of (81) is achieved. We introduce the following notation

$$q^{jr*} = \max_{k=1,2,\cdots,K} \gamma_k^{jr*}, \quad 1 \le r \le j \le J \tag{90}$$

$$q^{j*} = (q^{j1*}, q^{j2*}, \cdots, q^{jj*}) \tag{91}$$

$$w_i = \frac{1}{b_i} \sum_{r=1}^{j} (\gamma_j - q^{jr*}) b_r, \quad 1 \le i \le j \le J \tag{92}$$

$$\Gamma_i^{j*} = (\Gamma_i^{j1*}, \Gamma_i^{j2*}, \cdots, \Gamma_i^{jj*}) \tag{93}$$

$$= q^{j*} + w_i e_i^j, \quad 1 \le i \le j \le J \tag{94}$$

where e_i^j is a vector with j components, all of which are zero, except the ith one, which is equal to one. Note that the defintion of (94) implies that $\Gamma_i^{j*} \in \Delta_j$.

Theorem 5.1: The vector $(\gamma^{j1*}, \gamma^{j2*}, \cdots, \gamma^{jj*})$ is a convex combination of the j vectors $\{\Gamma_i^{j*}, i = 1, 2, \cdots, j\}, 1 \le j \le J$.

Proof: The theorem is obvious for $j = 1$ since in that case $\Delta_1 = \{(\gamma_1)\}$ is a singleton. Now consider the case where $2 \le j \le J$. We further assume that the vectors $\{\gamma_k^{j*} = (\gamma_k^{j1*}, \cdots, \gamma_k^{jj*}), k = 1, \cdots, K\}$ lie in the interior of Δ_j. (The theorem holds even when they are on the boundary, but the proof is more tedious.) Then, from (89), we have

$$g_k^{r*}(\gamma_k^{jr*}) = c_k^{j*}, \quad \text{for all } 1 \le r \le j. \tag{95}$$

Now consider a fixed r, $1 \le r \le j$. Since $g_k^{r*}(\eta)$ is a decreasing function of η, it follows that for all $x \in \Delta_j$ with $x_r \ge \gamma_k^{jr*}$ we have $g_k^{r*}(x_r) \le c_k^{j*}$ and there exists at least one $i \ne r$ such that $g_k^{i*}(x_i) \ge c_k^{j*}$. Hence

$$\max_{i=1,\cdots j} g_k^{i*}(x_i) = \max_{i=1,\cdots,j;i\ne r} g_k^{i*}(x_i). \tag{96}$$

Now, let $x \in \Delta_j$ be such that $x_r < \gamma_k^{jr*}$. Then, there exists a $y \in \Delta_j$ such that $x_r < y_r \le \gamma_k^{jr*}$ and $y_i < x_i$, for $i \ne r$. Hence we have

$$\max_{i=1,\cdots j} g_k^{i*}(x_i) = \max_{i=1,\cdots,j;i\ne r} g_k^{i*}(x_i) \tag{97}$$

$$\ge \max_{i=1,\cdots,j;i\ne r} g_k^{i*}(y_i) \tag{98}$$

$$= \max_{i=1,\cdots,j} g_k^{i*}(y_i). \tag{99}$$

Hence, for each $r, 1 \le r \le j$, and $x \in \Delta_j$ such that $x_r > q^{jr*}$ there is a $y \in \Delta_j$ such that $x_r > y_r \ge q^{jr*}$ and $y_i > x_i$, for $i \ne r$, and

$$\max_{i=1,\cdots,j} g_k^{i*}(x_i) \ge \max_{i=1,\cdots,j} g_k^{i*}(y_i) \tag{100}$$

for all $k = 1, \cdots, K$. This implies that $\gamma^{jr*} \le q^{jr*}$. The set $\{x \in \Delta_j : x_r \le q^{jr*}\}$ is the convex hull of the vectors $\{\Gamma_i^{j*}, i = 1, \cdots, j\}$. Hence the theorem follows. □

The above theorem can be used in two ways: First, it reduces the space over which one has to search for the system-optimal vector $(\gamma^{j1*}, \cdots, \gamma^{jj*})$. Secondly, it provides another candidate for an effective bandwidth vector with desirable properties, as described in the next section.

Additive Bandwidth Vectors

We begin by defining a class of effective bandwidth vectors. For $k = 1, 2, \cdots, K$ define

$$e_k^{j*} = \max_{r=1,\cdots,j} \{g_k^{r*}(\xi_{jr}^*)\}, \qquad \text{for } 1 \le j \le J \tag{101}$$

where $\{\xi_{jr}, 1 \le r \le j \le J\}$ is a given set of numbers so that $\xi_j = (\xi_{j1}, \cdots, \xi_{jj}) \in \Delta_j$. Define

$$e_k^* = (e_k^{1*}, e_k^{2*}, \cdots, e_k^{J*}). \tag{102}$$

We study e_k^* as a possible candidate for an effective bandwidth vector for the kth source.

Theorem 5.2:

$$(i)\ e_k^* \ge c_k^*, \quad \text{where } c_k^* \text{ is as defined in (89).} \tag{103}$$

$$(ii)\ \sum_{k=1}^{K} e_k^* \ge c^*(K) \text{where } c^*(K) \text{ is defined by (81).} \tag{104}$$

Proof: (i) Using (101) and (89) we get

$$e_k^{j*} = \max_{r=1,\cdots,j} \{g_k^{r*}(\xi_{jr}^*)\}, \tag{105}$$

$$\ge \min \max_{r=1,\cdots,j} \{g_k^{r*}(\gamma_{jr})\} \tag{106}$$

$$= c_k^{j*}. \tag{107}$$

The inequality follows because ξ_{jr}^* belongs to the set of γ_{jr} over which the minimum is taken. (ii) Using (101) and (89) we get

$$\sum_{k=1}^{K} e_k^{j*} = \sum_{k=1}^{K} \max_{r=1,\cdots,j} \{g_k^{r*}(\xi_{jr}^*)\} \tag{108}$$

$$\ge \max_{r=1,\cdots,j} \sum_{k=1}^{K} \{g_k^{r*}(\xi_{jr}^*)\} \tag{109}$$

$$\ge \min \max_{r=1,\cdots,j} \sum_{k=1}^{K} \{g_k^{r*}(\gamma_{jr})\}. \tag{110}$$

□

The property in (104) provides the justification for calling e_k^* an additive bandwidth vector.

Note that e_k^* is not an effective bandwidth vector of the usual type, i.e., it is not defined in terms of the characteristics of the kth source alone. In fact, it depends upon the source characteristics of all the sources in the system. However, it still provides an implementable method of call admission.

Next we consider two special sets of vectors $\{\xi_j, 1 \le j \le J\}$.

1) *Average of extreme points*: As a first choice, consider

$$\xi_j = \frac{1}{j} \sum_{i=1}^{j} \Gamma_i^{j*}, 1 \le j \le J \tag{111}$$

where Γ_i^{j*} is an extreme point of Δ_j as defined in (93). Clearly, $\xi_j \in \Delta_j$ for all $1 \le j \le J$.

2) *Average over sources:* As a second choice, consider

$$\xi_{jr} = \frac{1}{K} \sum_{k=1}^{K} \gamma_k^{jr*}, 1 \le r \le j \le J \tag{112}$$

where $\gamma_k^{j*} = (\gamma_k^{j1*}, \cdots, \gamma_k^{jj*})$ is the point where the minimum over the right-hand side of (89) is achieved. Thus the vector ξ_j is an average of the vectors γ_k^{j*} and hence is in Δ_j.

A Special Case

Here we consider a special case where each source produces a single-priority fluid. However, the total number of priorities is still J. The motivation is to study the multiplexing of sources of differing QoS requirements into a single buffer.

We say that a source is of type j if it produces traffic with QoS requirement ϵ_j (i.e., priority j traffic). Let K_j be the set of sources of type j. Now, consider a source $k \in K_t$. We have

$$\lambda_{k,i}^j = \begin{cases} 0, & \text{if } \ j \ne t \\ \lambda_{k,i}^j, & \text{if } \ j = t \end{cases} \tag{113}$$

and

$$\Lambda_{k,i}^j = \begin{cases} 0, & \text{for } \ t < j \le J \\ \lambda_{k,i}^t, & \text{for } \ 1 \le j \le t. \end{cases} \tag{114}$$

This, along with (25) implies that

$$A_k^j(\eta) = \begin{cases} A_k^t(\eta), & \text{for } \ 1 \le j \le t \\ -\frac{1}{\eta} G_k, & \text{for } \ t < j \le J \end{cases} \tag{115}$$

which further implies that

$$g_k^{j*}(\eta) = \begin{cases} g_k^{t*}(\eta), & \text{for } \ 1 \le j \le t \\ 0, & \text{for } \ t < j \le J. \end{cases} \tag{116}$$

Now consider the computation of the approximate bandwidth vector c_k^* as defined in (89). We have

$$c_k^{j*} = \min_{(\gamma_{j1},\cdots,\gamma_{jj}) \in \Delta_j} \max_{r=1,\cdots,j} \{g_k^{r*}(\gamma_{jr})\}, \qquad 1 \le j \le J \tag{117}$$

$$= \min_{(\gamma_{j1},\cdots,\gamma_{jj}) \in \Delta_j} \max_{r=1,\cdots,j} \{g_k^{t*}(\gamma_{jr})\}, \qquad 1 \le j \le t \tag{118}$$

$$= \min_{(\gamma_{j1},\cdots,\gamma_{jj}) \in \Delta_j} \max \{ \max_{r=1,\cdots,t} \{g_k^{t*}(\gamma_{jr})\}, 0\}, \quad t < j \le J. \tag{119}$$

The above minimum is achieved at $\gamma_{jr} = \gamma_k^{jr*}$ where

$$\gamma_k^{jr*} = \begin{cases} \gamma_j, & \text{for } 1 \le r \le j \le t \\ 0, & \text{for } 1 \le r \le j, t < j \le J \ . \end{cases} \tag{120}$$

Thus, computing the ξ_j vectors (using the average over sources method) for the additive bandwidth vectors using (25) is easy.

We get

$$\xi_{jr} = \frac{\sum_{t=1}^{r} |K_t|}{K} \gamma_j, \qquad 1 \le r \le j \le J. \tag{121}$$

Note that this method is particularly easy to use, since it involves using the fixed parameters γ_j and keeping track of $|K_t|$ for each $t = 1, 2, \cdots, J$.

VI. Computational Algorithm

We begin by describing a numerical algorithm to compute the minimum in (81) and (89). The algorithm uses ideas of steepest descent as well as of binary search, and is based upon the following simple theorem, which we state without proof.

Theorem 6.1: Let $h_i : (-\infty, 0] \to [0, \infty), 1 \le i \le j$ be bounded monotone decreasing functions. Let $(\eta_1, \eta_2, \cdots, \eta_j)$ be the value of $x = (x_1, x_2, \cdots, x_j)$ where the function $h(x) = \max_{i=1,2,\cdots,j} h_i(x_i)$ achieves its minimum over $x \in \Delta_j$. If, for a given $x \in \Delta_j$, $h_r(x_i) = \max_{i=1,\cdots,j}\{h_i(x_i)\}$ then, $\eta_r \ge x_r$.

This is a direct result of the structure of the minimization problem and the monotone nature of the h_i functions. It yields the following algorithm to find $(\eta_1, \eta_2, \cdots, \eta_j)$.

Algorithm A: **Given:** Monotone decreasing bounded functions $h_i : (-\infty, 0] \to [0, \infty)$. Two small positive numbers tolerance1 and tolerance2 to dictate the stopping criterion.

Aim: To find $(\eta_1, \eta_2, \cdots, \eta_j)$ as defined in Theorem 6.1.

Step 0: Set $L = \gamma_j \sum_{i=1}^{j} b_i(\frac{1}{b_1}, \frac{1}{b_2}, \cdots, \frac{1}{b_j})$.

Step 1: Compute

$$\delta = \frac{\gamma_j \sum_{i=1}^{j} b_i}{\sum_{i=1}^{j} L_i b_i}.$$

Set $x = \delta L$.

Step 2: Let

$$h_{\max} = \max_{i=1,2,\cdots,j} h_i(x_i)$$

and

$$h_{\min} = \min_{i=1,2,\cdots,j} h_i(x_i).$$

If $h_{\max} - h_{\min} \le$ tolerance1, stop. x is the desired η. If $-\max_{i=1,2,\cdots,j}\{x_i\} \le$ tolerance2, stop. x is the desired η.

Step 3: Set $L_r = x_r$ and go to step 1.

Remark: If the desired η vector is in the interior of Δ_j, the algorithm stops due to the first stopping condition in Step 2. If η is on the boundary of Δ_j , the algorithm stops due to the second stopping condition in Step 2.

Note that this algorithm can be used to compute c_k^{j*} in (89) as well as $c^{j*}(K)$ in (84) because both $g_k^{i*}(x_i)$ and $\sum_{k=1}^{K} g_k^{i*}(x_i)$ satisfy the properties of Theorem 6.1. Once the algorithm produces the (γ_k^{jr*}) vectors, they can be used to compute the additive bandwidth vector defined by (102). This algorithm is used in all the numerical work reported below.

VII. Numerical Results

Though most of this paper has concentrated on theoretical results, and the construction of approximate bandwidth vectors, we will show some numerical results. One can reasonably ask whether the procedures outlined above actually produce enough benefit to be worthwhile. In this short section, we give an example to show that priority differentiation may indeed be worthwhile. We use the same example as in [10]. Here we extend the voice source example of that paper to explore the effects of multiple priorities and multiple homogeneous sources.

Once again, we assume that the peak bit rate, including ATM overhead, is 72 170 b/s. The parameters of the two-state Markov chain controlling the source are: average talkspurt = 350 ms. and average silence duration = 650 ms. We used four different configurations:

1) *No priority differentation:* In this case, there is only one buffer threshold, 100 cells, which is the size of the buffer. The target loss probability used was the most stringent requirement of the traffic stream, 10^{-10}.
2) *Two priorities:* In this case, we used two buffer thresholds: 55 cells, for low-priority traffic and 100 cells, for high-priority traffic. The target loss probability for low-priority traffic is 10^{-1} and for high-priority traffic it is 10^{-10}. The peak bit rate is split evenly between the priorities.
3) *Three priorities:* For three priorities, we set the buffer thresholds at 55 cells (low priority), 91 cells (medium priority), and 100 cells (high priority). The target loss probabilities we set to 10^{-1} for low priority, 10^{-7} for medium priority, and 10^{-10} for high priority. We split the peak bit rate among the priorities as follows: 25% each for high and low priority traffic and 50% for medium priority.
4) *Four priorities:* Finally, for four priorities we set the buffer thresholds to 55 cells, 60 cells, 91 cells, and 100 cells (lowest priority to highest). We set the target loss probabilities to 10^{-1}, 10^{-4}, 10^{-7}, and 10^{-10} (again, lowest priority to highest). The peak bit rate was again split evenly among all the priorities.

In this experiment, we computed the additive effective bandwidth vector e_k^* defined in (102). Recall that the e_k^* vectors provide an upper bound for $c^*(K)$ (104) but that since all the sources in this case are identical, the system effective bandwidth vector and the additive effective bandwidth vectors are the same.

Table I shows the result of the computation of the effective bandwidth for the four different configurations described above.

The entries in Table I show the total effective bandwidth (in bits per second) required to support a given number of sources with a given number of priorities differentiated. We can look at the savings achieved from two different points of view: First is to look at the total effective bandwidth required by each of the priority configurations. Having two priorities saves about 8% of the bandwidth over no differentiation. Three priorities saves about 11% of the bandwidth and having four priorities

TABLE I
HOMOGENEOUS SOURCES: TOTAL BANDWIDTH USED

Sources	*Number of Priorities*			
	1	2	3	4
21	1,515,568	1,399,466	1,353,172	1,294,818
22	1,587,738	1,466,107	1,417,608	1,356,476
23	1,659,908	1,532,749	1,482,045	1,418,134
24	1,732,078	1,599,390	1,546,481	1,479,792
25	1,804,248	1,666,031	1,610,918	1,541,450

saves about 15% of the bandwidth in this case. Second, we can look at the savings from the point of view of the number of connections that would fit on a 1.544 Mb/s link. Note that we have ignored the usual framing and that also, ATM overhead prevents the usual 24 64kb/s connections from fitting in the DS1.

In this example, using four priorities, we can fit about 19% more connections into the given bandwidth than with no priority differentiation. With two and three priorities, we can fit 9.5% more connections.

In this case, we did not find optimal buffer thresholds, as we did in [10] and so the bandwidth savings are less than shown in that paper.

REFERENCES

[1] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "On the effectiveness of effective bandwidths for admission control in ATM networks," in *Proc. ITC-14*, 1994, pp. 411–420.

[2] I. Cidon, R. Guerin, and A. Khamisy, "On protective buffer policies," IBM Research Division, tech. rep. RC 18113, 1992.

[3] A. I. Elwalid and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic," *IEEE Trans. Commun.*, vol. 42, no. 11, pp. 2989–3002, 1994.

[4] ——, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Tran. Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[5] I. Cidon, R. Guèrin, and A. Khamisy, "A unified approach to bandwidth allocation and access control in fast packet-switched networks," in *Proc. INFOCOM-92*, 1992, pp. 1–12.

[6] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.

[7] L. Gun, A. Narayan, and V. G. Kulkarni, "Bandwidth allocation and access control in high speed networks," *Annals Op. Res.*, 1994.

[8] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1598–1608, 1988.

[9] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424–428, 1993.

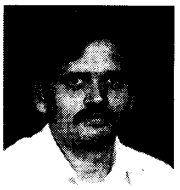
[10] V. G. Kulkarni, L. Gun, and P. F. Chimento, "Effective bandwidth vectors for two-priority ATM traffic," in *Proc. INFOCOM '94*, 1994.

[11] V. G. Kulkarni and T. E. Tedijanto, "Optimal cell discarding policy for color-coded voice traffic in ATM networks," IBM Corp., tech. rep. TR 29.1584, 1993.

[12] A. Y. M. Lin and J. A. Silvester, "Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system," *IEEE J. Select. Areas Commun.*, vol. 9, no. 9, pp. 1524–1536, 1991.

[13] K. Lindberger, "Analytical methods for traffic problems with statistical multiplexing in ATM networks," in *Proc. 13th Int. Teletraffic Cong.*, 1991.

[14] L. Tassiulas, Y. C. Hung, and S. S. Pannar, "Optimal buffer control during congestion in an ATM network node," in *Proc. 1992 Conf. Inform. Syst., Sci.*, 1992.



V. G. Kulkarni was born in Solapur, India, in 1955. He received the B.Tech. degree in mechnaical engineering from the Indian Institute of Technology, Bombay, in 1976, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1978 and 1980, respectively.

After a year at Georgia Institute of Technology as a Visiting Faculty, he moved to the University of North Carolina, Chapel Hill, in 1981, where he is currently a Professor in the Department of Operations Research. His main research interest is stochastic models. He has published numerous papers containing applications of stochastic processes to retrial queues, computer performance, fault tolerant systems, stochastic Petri nets, database systems, communications systems, effective bandwidths, stochastic fluid models, among others. He authored the graduate textbook *Modeling and Analysis of Stochastic Systems* (Chapman-Hall, 1995).

Dr. Kulkarni is an Associate Editor of *Operations Research Letters* and is a member of the editorial boasrd of *Stochastic Models.*

L. Gün (S'85–M'90–SM'93) was born in Brussels, Belgium, in July 1961. He received the B.A. degree in mathematics and the B.S. degree in electrical engineering from Bogazici University, Turkey, in 1983. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1986 and 1989, respectively.

He has held visiting positions at IBM, AT&T Bell Laboratories, INRIA, and George Mason University, Fairfax, VA. From August 1989 to January 1995, he was with the IBM High Bandwidth Architecture group, Research Triangle Park, NC, where he was the Lead Architect for bandwidth management and congestion control issues of IBM's Networking BroadBand Services (NBBS) architecture. From 1993 to January 1995, he was Manager of the NBBS Architecture and Emerging Technologies Departments. During 1993, he was an Adjunct Professor in the Operations Research Department, University of North Carolina, Chapel Hill. He is now with Motorola Information Systems Group, Mansfield, MA, where he heads the Networking Research Department. His current research interests are in ATM traffic management and quality of service issues, wireless ATM, and multimedia distribution over the cable plant. He has published numerous research papers in the areas of bandwidth management, congestion control, and dynamic routing issues for fast packet-switched networks, as well as computational probability, queueing theory, and stochastic control.

Dr. Gün served as the technical cochair and member of technical program committees of several international conferences. He is an active contributor to the traffic management and QoS groups in the ATM Forum.

P. F. Chimento was born in Monesson, PA, in 1950. He received the A.B. degree in philosophy from Kenyon College, Gambier, OH, in 1972, the M.S. degree in computer science from Michigan State University, East Lansing, in 1978, and the Ph.D. degree in computer science from Duke University, Durham, NC, in 1988.

He was with IBM Corporation from 1978 to 1994, holding various positions in design, development, testing, and architecture. Most recently, he was a member of the core team that developed IBM's Broadband Networking Services Architecture for high-speed packet and cell switching. In 1994, he took a leave of absence from IBM to accept a Visiting Faculty position at the University of Twente, The Netherlands, where he is a Member of the Centre for Telematics and Information Technology (CTIT) and the Tele-Informatics and Open Systems (TIOS) group, working on B-ISDN signaling and resource allocation issues and participating in Dutch and European telecommunications projects. He has published in the IEEE TRANSACTIONS ON COMPUTERS, *Operations Research,* and several conferences.

Dr. Chimento is a member of the ACM and ORSA (INFORMS).